387-396　　　　　　　　　　　　　　　　　　　　11106(16

动 物 学 研 究 1992，13(4):387—396　　　　ISSN 0254—5853
Zoological Research　　　　CN 53—1040/Q

削偶合今祖法的提出

李靖炎
（中国科学院昆明动物研究所　650107）

Q95/.3

摘要　以低等真核生物的5.8S rRNA为材料进行研究。发现在用今祖法消除了进化速度之差的影响所得到的差异矩阵d′中，各组理应彼此相等的数值会因核苷酸偶合的影响而成为不等值的。取各组中各自的的最大值作为受偶合的影响最小者，以之对今祖法所得到的分枝型式进行校正，即可得到更为正确的结果。在此基础上提出了对差异矩阵d′进行最大值成聚以代替原法中的平均值成聚的削偶合今祖法。新法在结果的正确性、稳定性和一致性上都优于原法。

关键词：分子进化树的构建，今祖法，同源大分子间的偶合，单细胞真核生物，5.8SrRNA

分子进化树，分子进化

引　言

不加权对群算术平均成聚（UPGMA法，简单成聚）本为构建表相图的方法，但自Nei(1975)指出可用以利用分子序列资料构建进化树以来，已广泛地得到了应用。但后来发现如果不同进化枝上的进化速度明显不等，用此法就可能得到错误的结果（Tateno，1978；Blanken等，1982）。Klotz与Blanken（1981）和W.-H. Li（1981）对此法作了巨大的改进，提出了今祖法以避免因进化速度之差所造成的错误。其基本原理在于利用参证，把不等速进化树的差异矩阵d转换成具有相同的分枝型式（topology）的等速进化树的差异矩阵d′，之后再依后者进行成聚（李靖炎，1988a）。

在实际应用今祖法时也发现一些问题。理论上，用任何物种的同源大分子作参证，只要此物种与建树物种不属于同一进化枝，就都应得到同一的结果。但实际上用不同的物种作参证，往往会得到彼此大不相同的结果。此外，矩阵d′既然代表等速进化树，其中就理应存在一系列的等值数值群，但实际上根本看不到。作者（1988a）的分析表明，上述两种矛盾都是由于同源大分子间的偶合造成的。理论上同源大分子间的差异程度应反映物种进化上的分歧程度，但核酸分子仅由4种核苷酸构成，在进化中会不断地发生核苷酸的偶合（平行替代与偶合替代），这些偶合会不等地降低同源大分子间的差异。即使进化速度完全一致，理应相等的差异值也会因此而变成不等。由此即可说明为何不同的参证会导致不同的结果。作者(1988b)曾推导出一个完整的公式，以求消除建树物种彼此间和参证物种与建树物种间同源大分子的偶合所造成的影响。但因当时无法估算建树物种彼此间的偶合值，被迫对此公式作了简化，以致只能消除不同参证所造成的误差。

本工作证明这一简化公式正是多物种联合参证的今祖法的理论基础。在力图削减建树物种同源分子间的偶合影响的基础上，作者对今祖法作了改进，提出了削偶合今祖法。其理论基础正是以前提出的未经简化的完整公式。

材 料

材料为涡鞭毛虫 *Crypthecodinium cohnii* 与 *Prorocentrum micans*、微胞子虫 *Vairimorpha necatrix*、贾第虫*Giardia lamblia*、纤毛虫 *Tetrahymena pyriformis*、酵母 *Saccharomyces cerevisiae*、锥体虫 *Crithidia fasciculata*与水稻 *Oriza sativa* 的5.8*S rRNA*，分别简称为 Cry、Pr、Va、Gi、Te、Sa、Cri与 Or。用为参证的是大肠杆菌、产甲烷菌*Methanobacterium thermoautotrophicum*与*Methanococcus vanielie*、嗜盐菌 *Halococcus morrhuae* 与 *Halobacterium halobium* 和依赖硫的嗜高温菌 *Thermoproteus tenax* 与*Desulfurococcus mobilis*的23S rRNA 分子中的 5.8S rRNA区，简称Ec、Mb、Mc、Hc、Hb、Tp与Dc。

同源大分子间的匹配往往有很大的主观性，特别是变异极大的段落。5.8S rRNA 5′端的头几个核苷酸与后 1/3 段即是如此。为尽力减少主观性，我们只利用保守性较强的2/3段来进行匹配（表 1）和计算差异值（表 2）。

表 1 7 种原核生物23S rRNA 分子中的5.8S rRNA 区与 8 种真核生物的 5.8S rRNA 的保守部分的匹配

Tab. 1 The matching of the conservative part of 5.8S rRNAs

Ec	UAAGCUGACA	CGGUGGAUGC	CCUGGCAGUC	AGAG-CGAUG	AAGGACGUGC	UAAUCUGCGA
Tp	CAAGCCGCC-	CGGUGGAUGG	CUCGGCUCGG	G-CGCCGAGG	AAGGGCGUGG	CAAGCUGCGA
Dc	GACGCCGCC-	CGGUGGAUGG	CUCGGCUCGG	G-CGCCGAGG	AAGGCCGUGG	CAAGCUGCGA
Hc	UAUGCCAAC-	UGGUGAAUAG	CUCGGCUCGA	GU-GCCGAUG	AAGGACGUGC	CAAGCUGCGA
Hb	UGUGCCACC-	UGGUGGAUAG	CUCGGCUCGG	AU-GCCGACG	AAGGACGUGC	CAAGCUGCGA
Mb	UAUGCCGUC-	UGGGGAAUGG	CUUGGCUUGA	GUCGCUGAUG	AAGGCCGUGG	CAAGCUGCGA
Mc	UACCCUACC-	UGGGGAAUGG	CUUGGCUUGA	AACGCCGAUG	AAGGACGUGG	UAAGCUGCGA
Pr	AACUUUCAG-	CGACGGAUGU	CUCGGCUCGA	-ACAACGAUG	AAGGGCGCAG	CGAAAUGUGA
Cry	AACUUUCAG-	CAGUUGAUUC	CUUGG-UUCA	GACCUCGAUG	AAGGGCGCUG	CGAAA-GUGA
Sa	AACUUUCAA-	CAACGGAUCU	CUUGG-UUCU	CGCAUCGAUG	AAGAACGCAG	CGAAAUGCGA
Cri	AACGUGUCG-	CGAUGGAUGA	CUUGGCUUCC	UAUCUCGUUG	AAGAACGCAG	UAAAGUGCGA
Or	GACUCUCGG-	CAACGGAUAU	CUCGGCU-CU	CGCAUCGAUG	AAGAACGUAG	CGAAAUGCGA
Te	AACUUUCAA-	CGGUGGAUAU	CUUGGUUCCC	GUGA-CGAUG	AAGAACGCAG	CGAAAUGCGA
Gi	AACGCCCCGC	CGGCGGAUGC	CUCGGC-CCG	GGCGGCGACG	AAGAGCGCGG	CGGAGCGCGA
Va	???ACCCACA	CAUGGGAUCA	AUAGGAUACC	-AUAACGAUG	AAGGUCGUAA	UAGAAUACGA

Ec	UAAGCGU-GG	UAAUGAUAUG	ACA-CCGUUA	UAACC--GGC	G-AUCUCCUA	AUG
Tp	UAAGCCCGGG	GUAGCCGCAA	GCGGGCGUU-	GAACC--CGG	G-AUUCCCGA	AUG
Dc	UAAGCCCGGG	GUAGGCGCAG	GCAGCCGUU-	GAACC--CGG	G-AUCGCCGA	AUG
Hc	UAAGCUCAGG	GGAGCCGCAC	GGAGGCGAA-	GAACC--UGA	G-AUUUCCGA	AUG
Hb	UAAGCCUUAG	GGAGCCGCAU	GCAUGCUAA-	GAACU--CAG	G-AUCUCCUA	AUG
Mb	UAAGCCCAGG	GGAGGAGCAU	GCAUCCUUG-	GAACC--UGG	G-AUUGCCGA	AUG
Mc	UAAGCCUAGG	GGAGGCGCAU	ACAGCCUUU-	GAACC--UAG	G-AUUUCCGA	AUG
Pr	UAAGCAAUGU	GAA-UUGCAG	AAUUCCG-U-	GAACCAAUAG	GGACUU--GA	ACG
Cry	AUGGCA-UGU	GAA--UGCAG	GCAUCCG-G-	GAAUUGAGAG	CUUCUU--GA	AUG
Sa	UACGUAAUGU	GAA-UUGCAG	AAUUCCG-G-	GAAUCAUCGA	AU-CUUU-GA	ACG
Cri	UAAGUGGUAU	GAA-UUGCAG	AAU-CAU-U-	CAAUUACCGA	AU-CUUU-GA	ACG
Or	UACCUGGUGU	GAA-UUGCAG	AAUCCCG-U-	GAACCAUCGA	GU-CUUU-GA	ACG
Te	UACGUAAUGC	GAA-UUGCAG	AA--CCG-C-	GAGUCAACAG	AU-CUUU-GA	AAG
Gi	GACGCGGUGC	GGACCCGCCC	GC-CCCG-A-	GAAGCACCGA	CC-CUC--GA	ACG
Va	-AAGUA-UAU	-UA-UU----	UACC--U-	GA-UUAAU--	AUA-UU	

表2　依 据 表 1 得 到 的 差 异 矩 阵 d

Tab. 2　Difference matrix d of 5.8S rRNAs

d	Sa	Or	Te	Cri	Pr	Cry	Va	Gi
Ec	58.41	53.98	55.86	51.79	53.1	58.41	61.165	56.64
Tp	54.95	48.65	51.35	53.15	43.24	52.25	61.76	40.18
Dc	53.15	45.95	49.55	51.35	42.34	50.45	61.76	41.95
Hb	60.36	54.05	53.15	55.86	50.45	57.56	62.745	48.21
Hc	54.95	49.56	50.45	55.86	45.95	54.05	61.76	45.95
Mb	54.95	53.15	54.05	54.95	46.85	51.35	62.745	50.00
Mc	52.25	48.65	50.45	48.66	40.54	49.55	58.82	51.79
Sa	—	15.89	19.63	28.97	25.00	33.33	49.00	43.12
Or		—	28.04	30.84	26.85	39.81	51.00	38.53
Te			—	33.02	26.85	33.33	50.00	43.52
Cri				—	36.11	43.52	47.00	44.95
Pr					—	29.91	47.96	42.20
Cry						—	55.56	45.87
Va							—	65.35
Gi								—

方 法 与 结 果

一、UPGMA法　对 8 种真核生物的5.8S rRNA 的差异值d 作简单成聚，得到图1，其中显然含有错误，例如在涡鞭毛虫与锥体虫的关系问题上。

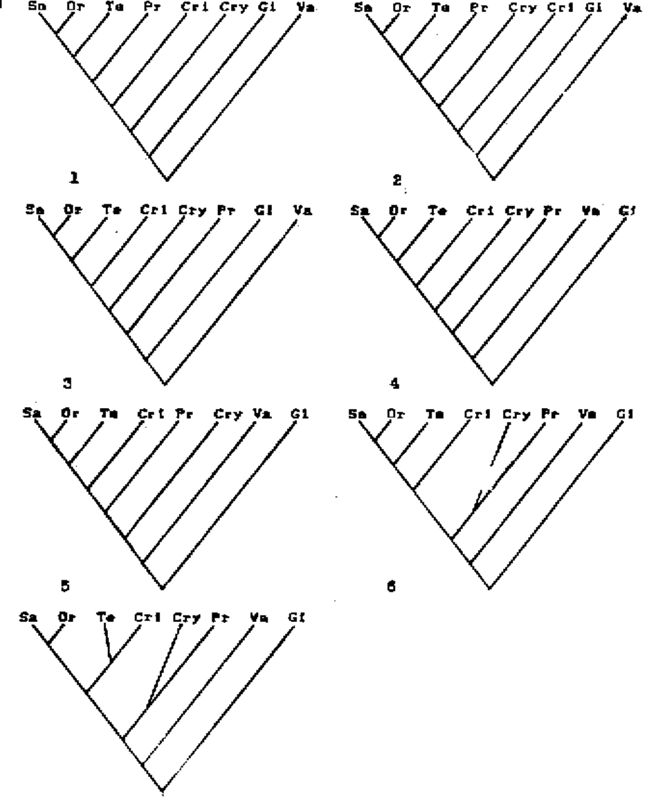


图 1 — 7　依不同方法所得到的分枝型式

Figs. 1 — 7　Topologies obtained with various methods

1. UPGM法所得
2. 今祖法，以大肠杆菌Ec作为参证所得
3. 今祖法，以产甲烷球菌Mc作为参证所得
4. 今祖法，以热变形菌Tp作为参证所得
5. 今祖法，以产甲烷杆菌Mb或脱硫球菌 Dc 作为参证所得；以 7 种原核生物联合参证所得也为此图
6. 今祖法，以盐杆菌Hb或盐球菌Hc作为参证所得。依削偶合今祖法，无论以Ec、Tp、Hb、Mb 或Mc为参证，所得皆为此图；以 7 种原核生物或 7 种中的任何 6 种作联合参证，所得也为此图
7. 削偶合今祖法，以盐球菌 Hc 或脱硫球菌Dc为参证时所得

二、今祖法 利用 $d'i_ai_b = di_ai_b - ri_aj - ri_bj$ 的公式可把差异矩阵 d 转换为代表等速进化树的矩阵 d'。i 代表建树物种的同源大分子，j 为参证物种的同源分子，rij 代表以 j 为参证时 i 的校正值，其基本形式即 dij。对诸 rij 可共同加减任一个值，而不影响最后的结果。对整个矩阵也是如此。

理论上用不同的细菌同源分子作参证应得到同一的结果。但分别以 7 种细菌的5.8 S rRNA 区作参证，却得到 5 种不同的结果（图 2 — 6）。以两种涡鞭毛虫应构成一个单系类群，而非并系类群来判断，只有图 6 可能正确。但即使以 7 个种作联合参证（表 3），也只能得到图 5 。可见今祖法即使作联合参证也并不一定能得到正确的结果。

表 3 以 Ec、Tp、Dc、Hb、Hc、Mb、Mc 等 7 种原核生物联合参证从表 1 得到的差异矩阵d'

Tab. 3 Difference matrix d' (reference, Ec, Tp, Dc, Hb, Mb & Mc)

d'	Or	Te	Cri	Pr	Cry	Va	Gi
Sa	15.89	18.0754	26.4514	29.5014	30.51	38.0324	45.8714
Or	—	31.491	33.327	36.357	41.9956	45.038	46.287
Te		—	33.9524	34.8024	33.961	42.4834	49.7224
Cri			—	43.0984	43.187	38.5194	50.1884
Pr				—	36.597	46.4994	54.4584
Cry					—	46.778	50.807
Va						—	62.1394
Gi							

$d'i_ai_b = di_ai_b - ri_a - ri_b + 14.0084$

rSa = 9.507 rOr = 4.5014 rTe = 6.056 rCri = 7.02

rPr = 0 rCry = 7.3214 rVa = 15.469 rGi = 1.76

三、受偶合影响最小的 d' 值及其利用 既然矩阵 d' 在理论上代表等速进化树，其中理应存在一系列的等值数值群。例如在据以得到图 5 的矩阵 d'（表 3）中，d'SaGi、d'OrGi、d'TeGi、d'CriGi、d'PrGi、d'CryGi、d'VaGi 等就应是等值的。但实际上并不如此。这可以认为是由于不同程度的偶合使它们不等地减小所致；其中的最大值则可以视为受偶合影响最小的一个。

各理论等值数值群中的最大值不仅可用以衡量各群中其他数值受偶合影响的相对大小，还可借以判断哪种分枝型式更为正确。以图 5 和图 6 为例。在这两种进化树中，如无偶合的干扰都应存在如下的等式关系：d'SaTe = d'OrTe = A,

d'SaCri = d'OrCri = d'TeCri = B,

d'SaPr = d'OrPr = d'TePr = d'CriPr = C,

d'SaCry = d'OrCry = d'TeCry = d'CriCry = D,

d'SaVa = d'OrVa = d'TeVa = d'CriVa = d'PrVa = d'CryVa = E,

d'SaGi = d'OrGi = d'TeGi = d'CriGi = d'PrGi = d'CryGi = d'VaGi = F.

如图5正确，则 A＜B＜C＜D＜E＜F，而且 d'PrCry 等于 D；如图6正确，则 A＜B＜C = D＜E＜F，而且 d'PrCry 小于 D。虽则由于偶合的干扰，上述的等式关系实际上并不

存在，但上述的不等式关系却应能通过各理论等值数值群中的最大值而表现出来。从表3中各个群的最大值 d'SaTe、d'TeCri、d'CriPr、d'CriCry、d'CryVa、d'VaGi和d'PrCry 的具体数值看，显然正确的是图6，而非图5。用这种方法已经证明了图1—4及其他一些进化树全都是错误的。

四、削偶合今祖法 上述的结果提示我们，可以对矩阵 d' 中的数值不是依最小的平均值，而是依最小的最大值来进行成聚。这就是削偶合今祖法。

仍以表3中的矩阵 d' 为例。按原今祖法只能得到图5，而按照削偶合今祖法则可得到图6。关键在于依原法取平均值，d'(SaOrTeCri) Pr = 35.9398小于 d'PrCry (36.597)，因此只能得到图5；而按新法取最大值，d'(SaOrTeCri) Pr = 43.0984，显著地大于d'PrCry，所得的因此是图6。

以7物种分别参证，根据所得的矩阵 d' 按原法成聚时，如前所述全得到5种不同的分枝型式；然而如按新法进行成聚，无论是以Ec、Mc、Mb、Hb 或Tp 作为参证，都会得到图6，而以Hc或 Dc 为参证则会得到图7。在图7中两种涡鞭毛虫也是构成单系，但其中纤毛虫与锥体虫的关系是难以置信的。

以数个物种作联合参证并依削偶合今祖法建树，结果更为明显。以7物种中的任意5种联合参证，在全部21种可能的组合中，只有一种会得到图7，其余所得的全是图6。以任意6种或全部7种联合参证，则都只会得到图6。

削偶合今祖法现已可在微机上自动进行。

讨　论

一、今祖法的意义及其改进 UPGMA法迄今是构建分子进化树的主要方法之一，但有两个因素往往妨碍它取得正确的结果：1.进化速度的差异，2.同源大分子间的偶合。今祖法的意义即在于避免了前一因素的干扰，但由于不同参证物种与各建树物种同

表4 以大肠杆菌Ec为参证从表1得到的差异矩阵d'

Tab.4 Difference matrix d' (reference: Ec)

d'	Sa	Or	Te	Cri	Pr	Cry	Va	Gi
Sa	—	15.89	17.75	31.16	25.88	28.9	41.815	40.46
Or		—	30.59	37.46	32.16	39.81	48.245	40.3
Te			—	37.76	30.28	31.45	45.365	43.41
Cri				—	43.61	45.71	46.435	48.91
Pr					—	30.79	46.085	44.85
Cry						—	48.375	43.21
Va							—	59.935
Gi								—

$d'i_a i_b = d'i_a i_b - ri_a Ec - ri_b Ec + 8.81$

rSaEc = 6.62　rOrEc = 2.19　rTeEc = 4.07　rCriEc = 0

rPrEc = 1.31　rCryEc = 6.62　rVaEc = 9.375　rGiEc = 4.85

表5　以大肠杆菌Ec为参证得到的差异矩阵DC′(d′值据表4)

Tab. 5　Difference matrix DC′(reference: Ec)

DC′	Or	Te	Cri	Pr	Cry	Va	Gi
Sa	15.89	18.0754	26.4514	29.5014	30.51	38.0324	45.8714
Or	—	31.491	33.227	36.357	41.9956	45.038	46.287
Te		—	33.9524	34.8024	33.961	42.4834	49.7224
Cri			—	43.0984	43.187	38.5194	50.1884
Pr				—	36.597	46.4994	54.4584
Cry					—	46.778	50.807
Va						—	62.1394
Gi							—

$DC'_{i_a i_b} = d'_{i_a i_b} - CO'_{i_a Ec} - CO'_{i_b Ec} + 13.8624$

$d'_{i_a i_b}$ are shown in table 2.

CO′SaEc = 7.219　CO′OrEc = 6.6434　CO′TeEc = 6.318

CO′CriEc = 11.352　CO′PrEc = 3.022　CO′CryEc = 5.0334

CO′VaEc = 10.426　CO′GiEc = 1.232

源大分子间有程度各不相同的偶合，结果就造成了不同的矩阵 d′ 和不同的分枝型式。作者（1988 b）曾通过一系列的数学推导得到了"$DO'_{i_a i_b}=d'_{i_a i_b}+CO'_{i_a i_b}-CO'_{i_a j}-CO'_{i_b j}$"的公式，$DO'_{i_a i_b}$代表已消除了各建树物种同源大分子间的偶合影响$CO'_{i_a i_b}$和参证物种与建树物种分子间的这种影响$CO'_{ij}$的$d'_{i_a i_b}$值。但因当时尚未找到估算$CO'_{i_a i_b}$值的方法，被迫对此公式作了明知不妥的简化，即假定诸建树物种的同源大分子间的偶合程度都大致相似，因而可予忽略，从而提出了如下的简化公式：$DC'_{i_a i_b}=d'_{i_a i_b}-CO'_{i_a j}-CO'_{i_b j}$。这一简化公式实际上正是多物种联合参证的今祖法的理论基础。比较表5与表3即可清楚地看到这一点。此法的缺陷看来也正在于上述的简化。

通过本文所述的工作，现已能对大部分$CO'_{i_a i_b}$的大小进行估算，从而使$DO'_{i_a i_b}$的公式得到应用。应用结果证明实际也就是联合参证的削偶合今祖法（$d'_{i_a i_b}$加上$CO'_{i_a i_b}$就成了最大值）。

二、削偶合今祖法的优越性　1.今祖法会因参证物种的不同而得到众多各不相同的结果。然而利用同样一些单个的参证物种，削偶合今祖法在绝大多数情况下却可得到一致并且正确的结果。

2.利用削偶合今祖法，作者得到了国际上迄今利用低等真核生物的 5.8S rRNA 建立分子进化树所得到的最好的结果。5.8S rRNA 是一种仅含150多个核苷酸的小分子rRNA，所含信息量仅及大分子rRNA所含的几十分之一。但我们用削偶合今祖法在5.8S rRNA上所得到的结果，与Sogin等（1989）、Perasso等（1989）和Lenaers等（1989）在大分子rRNA上所得到的结果基本上一致。这样好的结果是Maroteaux等（1985）、Walker（1985）过去在5.8S rRNA上都未曾得到过的。

我们的结果与别人用大分子rRNA为材料所得结果的唯一不同是在纤毛虫与涡鞭毛虫的亲缘关系问题上。从5.8S rRNA上确实看不出这两者有什么紧密关系。用其他建

树方法也看不出。但应指出，认为这两者有紧密的亲缘关系的看法迄今还缺乏细胞生物学和其他分子生物学的旁证。根据本文的结果，涡鞭毛虫类是介于具有70 S 型核糖体的微胞子虫、贾第虫与具有80 S 型核糖体的纤毛虫、锥体虫之间。这与涡鞭毛虫具有75 S 的核糖体的报道（Steele，1980）相吻合。

作者曾以本文中所用的 8 种真核生物的 5.8S rRNA 作为参证，用削偶合今祖法研究本文中所用的 7 种原核生物的进化关系。所得结果与当前国际上公认的看法是一致的，即原核生物首先区分为原细菌与真细菌两大枝，前者再分歧为依赖于硫的嗜高温的一枝与包括产甲烷菌与嗜盐菌的一枝。这一结果也为Gouyl 与 W.-H. Li（1989）用别的方法在大分子rRNA上得到。

三、各物种同源大分子间偶合影响的估算　本工作可以估算同源大分子间的偶合程度，因为$CO'i_ai_b$即$DO'i_ai_b$与$DC'i_ai_b$之差。DCi_ai_b可依公式求得，也可用联合参证时所得的$d'i_ai_b$来代替。在一个由多个 DC' 值所构成的理论等值数值群中，数值个数越多，其中的最大值就越可能近似于DO' 值。在实际应用上，可用作联合参证所得的矩阵d'中各理论等值数值群中的最大值作为DO'的近似值，而以其与各d'值之差作为各 CO' 的近似值。例如从表 5 中可以看出，贾第虫与酵母 5.8S rRNA 间的偶合程度即显然大于贾第虫与纤毛虫之间的，而后者又远大于贾第虫与微胞子虫之间的。

理论等值数值群中的个数越少，其最大值就越可能距DO' 值较远。但即使只有两个数值，在进化树的构建上也远比只有单一数值为优。图 7 中的错误与 d'TeCri 是一不成群的单一数值直接相关。如果它是两个或三个理论等值的数值中的一个，从中取得最大值来成聚，图 7 可能就不会得到。

致谢　本文曾蒙美国Texas大学种群统计与群体遗传研究中心 Wen-Hiung Li 教授提出宝贵意见。

参　考　文　献

李靖炎. 1988 a. 分子进化研究中的今祖法，其理论基础、存在问题和解释. 动物学研究，9(2):141—150.

李靖炎. 1988 b. 一种考虑到不同物种同源大分子间的偶合关系的新的今祖法. 动物学研究，9(4):327—334.

Blanken, R. L., L.C. Klotz, A.G. Hinnebusch. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. *J. Mol. Evol.*, 19: 9-19.

Gouyi, M. and W.-H. Li. 1989. Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. Nature, 339 (6220): 145-147.

Klotz, L. C. and R. L. Blanken. 1981. A Practical method for calculating evolutionary trees from sequence data. *J. Theor. Biol.*, 91:261-272.

Lenaers, G., L. Maroteaux, B. Michot, *et al.* 1989. Dinoflagellates in evolution. A molecular phylogenetic analysis of large subunit rRNA. *J. Mol. Evol.*, 29: 40-51.

Li, W.-H. 1981. A simple method for constructing phylogenetic trees from distance matrices. *Proc. Natl. Acad. Sci. USA*, 78: 1085-1089.

Maroteaux, L., M. Herzog and M.-O. Soyer-Gobillard. 1985. Molecular organization of dinoflagellate rRNA: evolutionary implications of the deduced 5.8S rRNA secondary structures. *BioSystems*, 18:307-319.

Nei, M. 1975. Molecular population genetics and evolution. North-Holland, Amsterdam.

Perasso, R., A. Baroin, J. H. Qu *et al.* 1989. Origin of algae. *Nature*, 339: 142-144.

Sogin, M.L. *et al*. 1989. Phylogenetic meaning of the kingdom concept: an unusual rRNA from Giardia lamblia. *Science*, 243: 75—77.
Steele, R. F. 1980. Ph. D. thesis, Yale University.
Walker, W. 1985. 5S and 5.8S rRNA sequences and protist phylogenetics. *BioSystems*, 18: 269—278.

THE PDARIC METHOD FOR CONSTRUCTING MOLECULAR EVOLUTIONARY TRESE FROM SEQUENCES DATA

Li Jingyan
(Kunming Institute of Zoology, Academia Sinica 650107)

The sequence data used were the conservative part of 5.8S rRNAs of seven species of protists (yeast, ciliate, trypanosomatid, microspora, diplomonad and two species of dinoflagellates) and of one species of higher plant (rice). The corresponding part in 23S rRNAs of seven species of prokaryotes (eubacterium, halophilic archaebacteria, methane-producing archaebacteria and sulfur-dependent thermophilic archaebacteria) were separately or integrately used as references.

The evolutionary tree constructed by UPGMA method (fig. 1) is obviously incorrect. This fact probably means that the evolutionary rates in different branches of the phylogenetic tree are quite different.

The present-day ancestor (PDA) method can, in theory, transform the difference matrix (matrix d) of a phylogenetic tree with unequal evolutionary rates into the matrix d' representing a tree with the same topology, but with equal rates. Then, from the matrix d' the real topology will be easily obtained. However, beyond expectation, from the same matrix d' of 5.8S rRNAs (table 2) five different topologies (fig. 2—fig. 6) were obtained by this method when seven reference species were separately used. These results contradict the theory. When these trees were judged by the criterion that the two species of dinoflagellates should construct a monophyletic group, only one among them (fig. 6) might be correct. Even when all the seven reference species were used integrately, the tree obtained (fig. 5) was not correct. Therefore, even when lots of species are used as an integrated reference group, PDA method still can not guarantee to obtain the correct topology.

Another contradiction about this method is as the following. In theory, since matrix d' represents a tree with constant rate, there must exist a series

of groups of d' values which are equal to each other within each group. Nevertheless, these theoretically widely existing equal-valued elements can not be found in any matrix d'.

Both contradictions were explained by the author (1988a) to be produced by various degrees of coincidences among pairs of homologous macromolecules. The originally equal difference values will become unequal when these values are unequally reduced because of coincidences. Therefore, the largest d' value of a group can be seemed as the value reduced at the least extent in the whole group.

In present work the author found that these largest d' values from respective groups could be used to represent these groups to verify the correctness of the topology obtained. With these largest values the author proved that all the topologies in fig. 2 to fig. 5 are wrong and only the one in fig. 6 is correct.

The successful results described above enlightened us that we'd better to make the clustering with the largest d' values, rather than with the mean values as in original PDA method. The new method was called PDARIC method which means the present-day ancestor method with reduction of the influences of coincidences.

When the same seven reference species were separately used, from the same matrix d only two topologies were obtained by the new method. Five separate reference species made the topology in fig. 6 and the other two made that in fig. 7. Although the two species of dinoflagellates also construct a monophyletic group in fig. 7 the topology seems to be incorrect, because ciliate and trypanosomatid also construct a similar group in this tree. When arbitrary five from the seven reference species were used integrately, among all 21 possible reference combinations, 20 got topology in fig. 6 and only one got that in fig. 7. If arbitrary six or all seven reference species are used as integrated reference groups, only the topology in fig. 6 can be obtained.

The previous work of the author (1988b) deduced a formula DO':

$$DO'_{i_a i_b} = d'_{i_a i_b} + CO'_{i_a i_b} - CO'_{i_a j} - CO'_{i_b j}.$$

$DO'_{i_a i_b}$ means the difference value between the two homologous macromolecules of studied species (i) which has removed the influence of the coincidence between the two macromolecules of studied species ($CO'_{i_a i_b}$) and those (CO'_{ij}) between the molecules of studied species and reference species (j). From matrix DO' the real topology should be obtained. However, the author did not find the way to estimate $CO'_{i_a i_b}$ value at that time. Therefore, under a very questionable assumption that all the influences of coincidences among pairs

of i are alike, the formula was simplified into formula DC':

$$DC'i_ai_b = d'i_ai_b - CO'i_aj - CO'i_bj.$$

In present work, the formula DC' is found to be the base of the original PDA method with multiple reference group. For example, the matrix d' obtained with seven reference species used integrately (table 3) is totally the same as the matrix DC' (table 5). Besides, the formula DO' is found to be the base of PDARIC method in theory.

Key words: Constructing molecular evolutionary tree, Present-day ancestor (PDA) method, Coincidences between homologous macromolecules, Unicellular eukaryotes, 5.8S rRNA